Insertions and deletions in the great tit genome

Henry Barton

01/05/18

Introduction

Insertions and deletions

- short INDELs: sections of DNA $<$ 50bp that are deleted or inserted in a genome
- deletion bias in most organisms
- deletions more deleterious than insertions

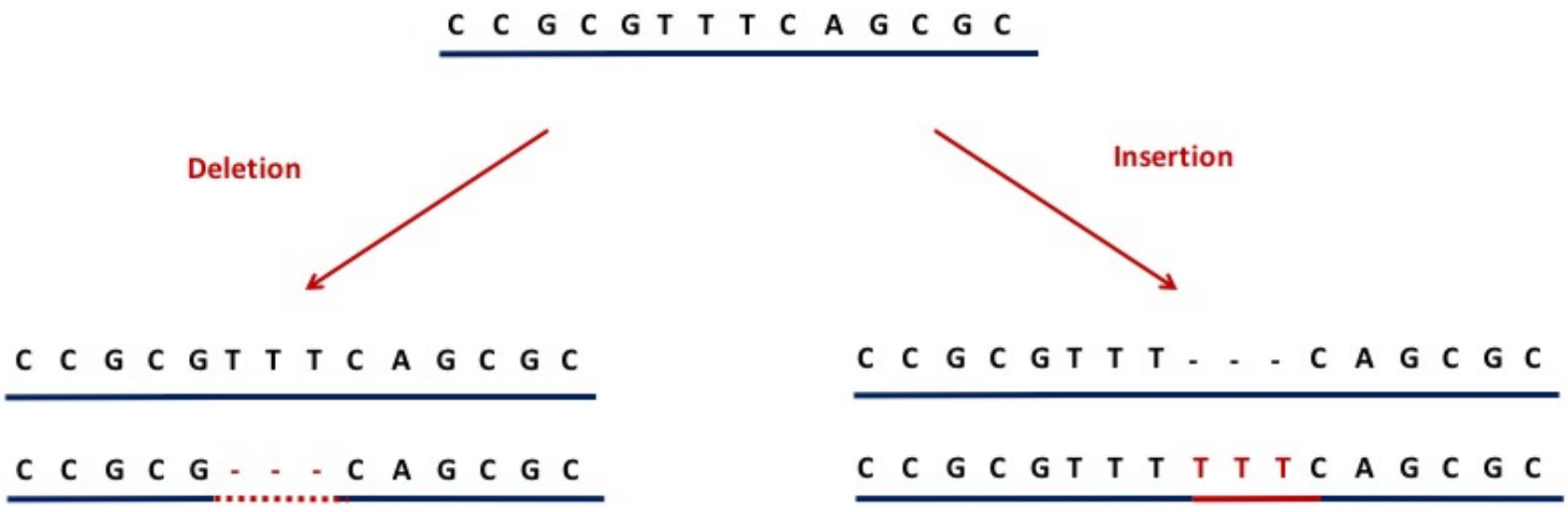


Figure 1: indel_diag

INDELs often overlooked

- Disproportionately occur in repetitive sequence
- Hard to align
- Often occur in hotspots
- 1/8 as frequent as SNPs in humans

(Earl et al., 2014; Montgomery et al., 2013)

The importance of INDELs in genome evolution

- Influence genome size:
  - low deletion rate $\rightarrow$ large genomes?
  - high deletion rate $\rightarrow$ compact genomes?

(Leushkin and Bazykin, 2013; Nam and Ellegren, 2012; Ometto et al., 2005; Sun et al., 2012)

The importance of INDELs in genome evolution

- Influence genome size:
  - low deletion rate $\rightarrow$ large genomes?
  - high deletion rate $\rightarrow$ compact genomes?

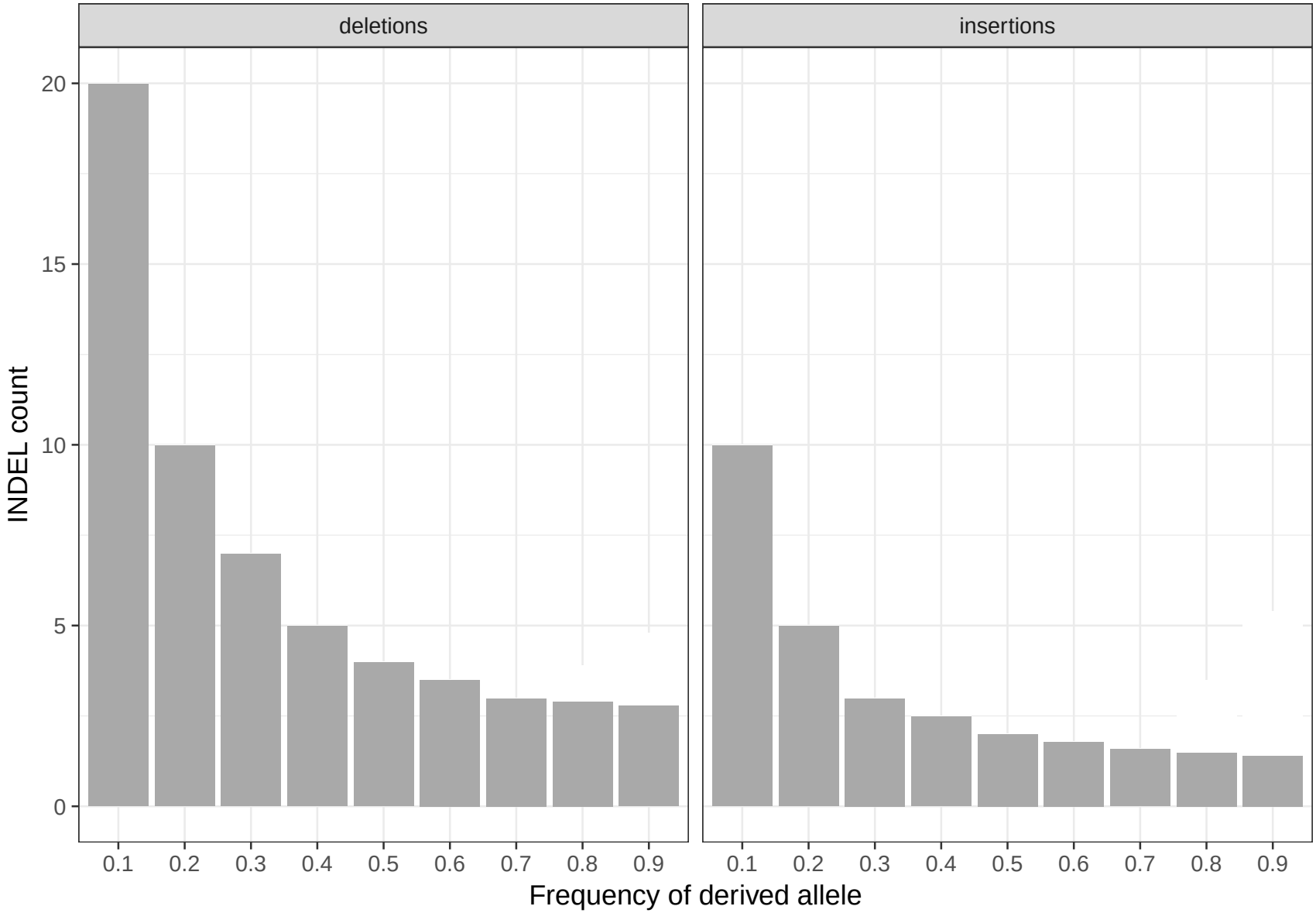
(Leushkin and Bazykin, 2013; Nam and Ellegren, 2012; Ometto et al., 2005; Sun et al., 2012)

INDEL mutation

INDEL selection

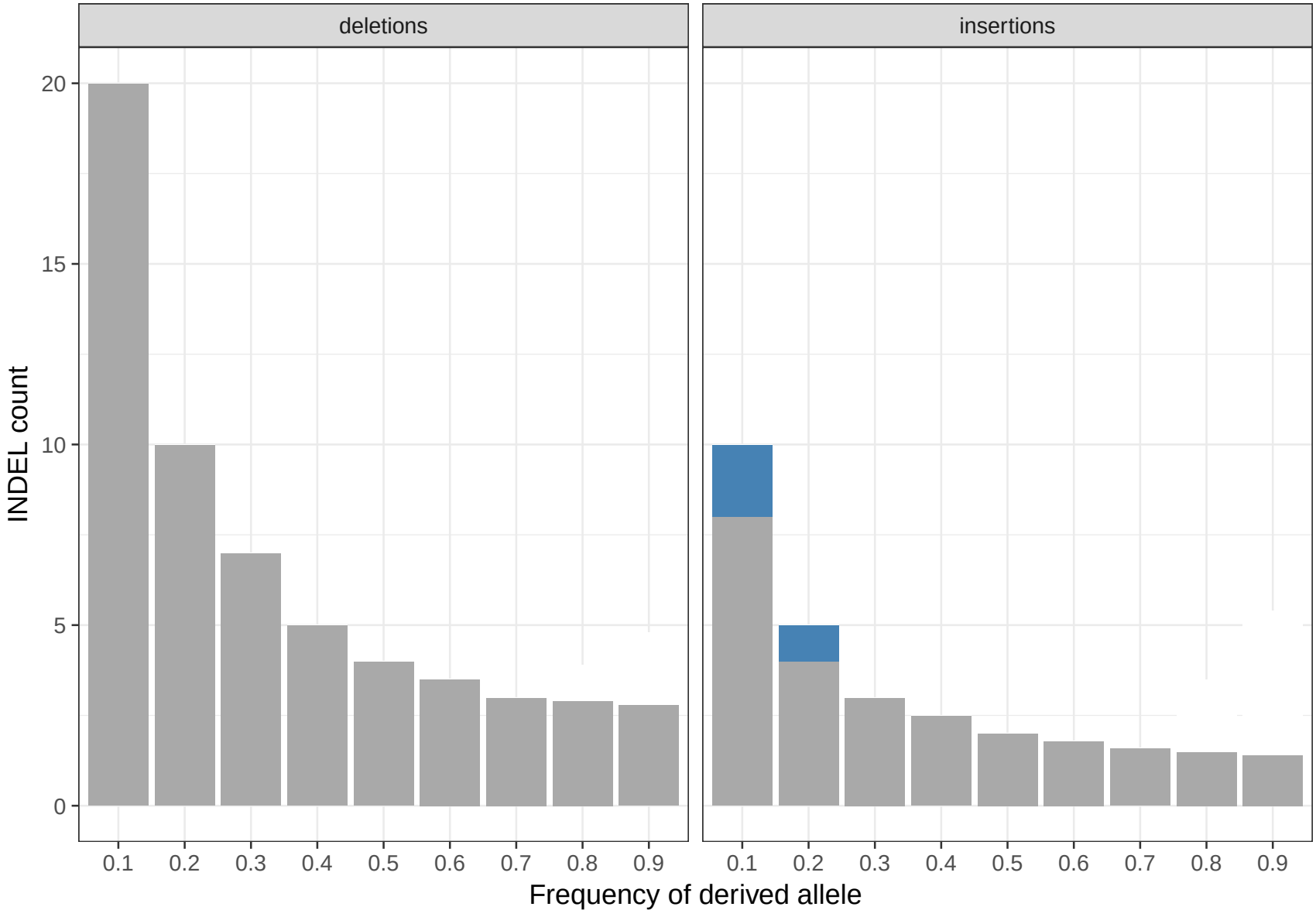
- Deletions
  -
- Insertions may be favoured:
  - biased gene conversion
  - minimum intron size
  - polarisation error

Importance of polarisation error



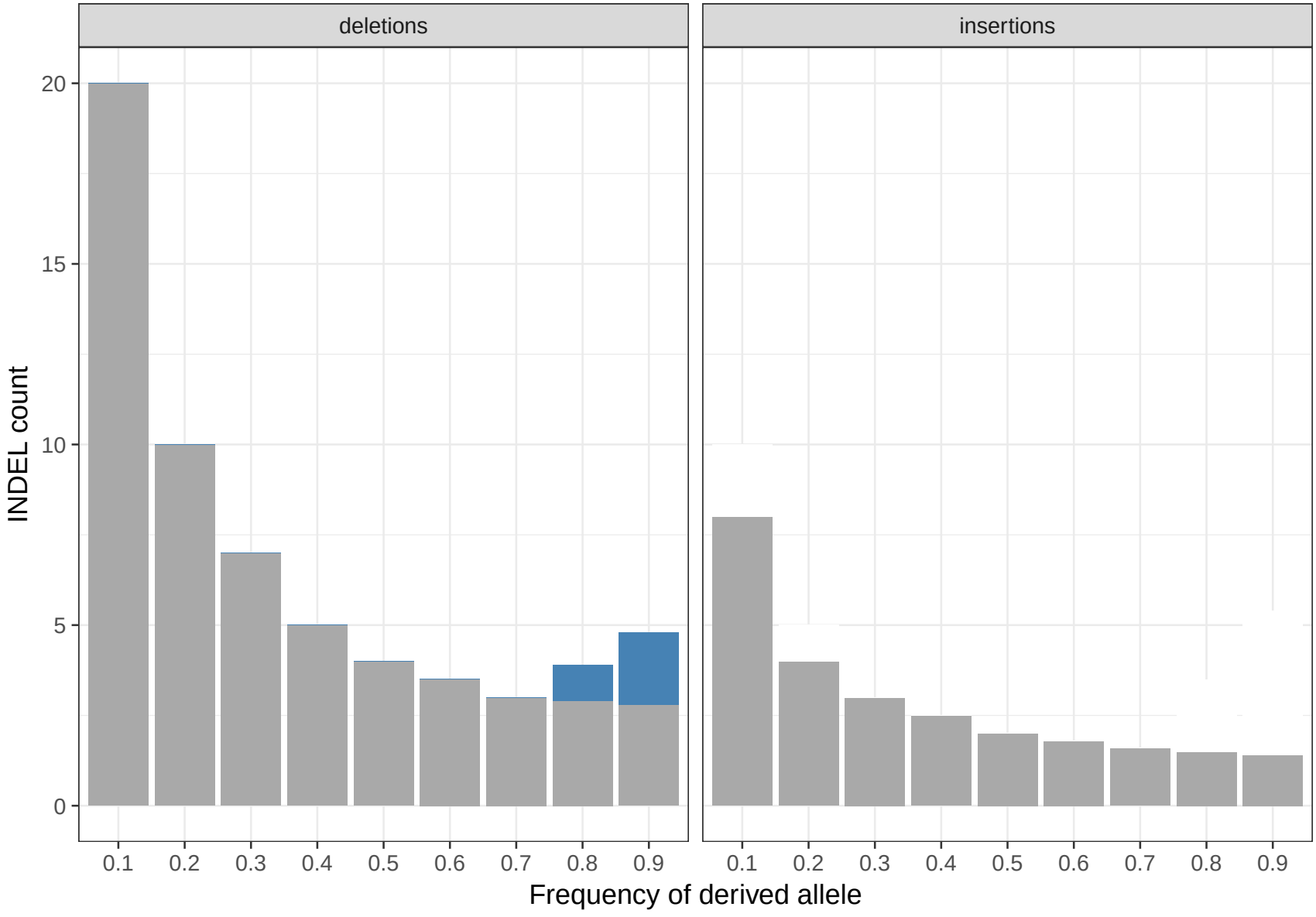
(see Hernandez et al., 2007)

Importance of polarisation error



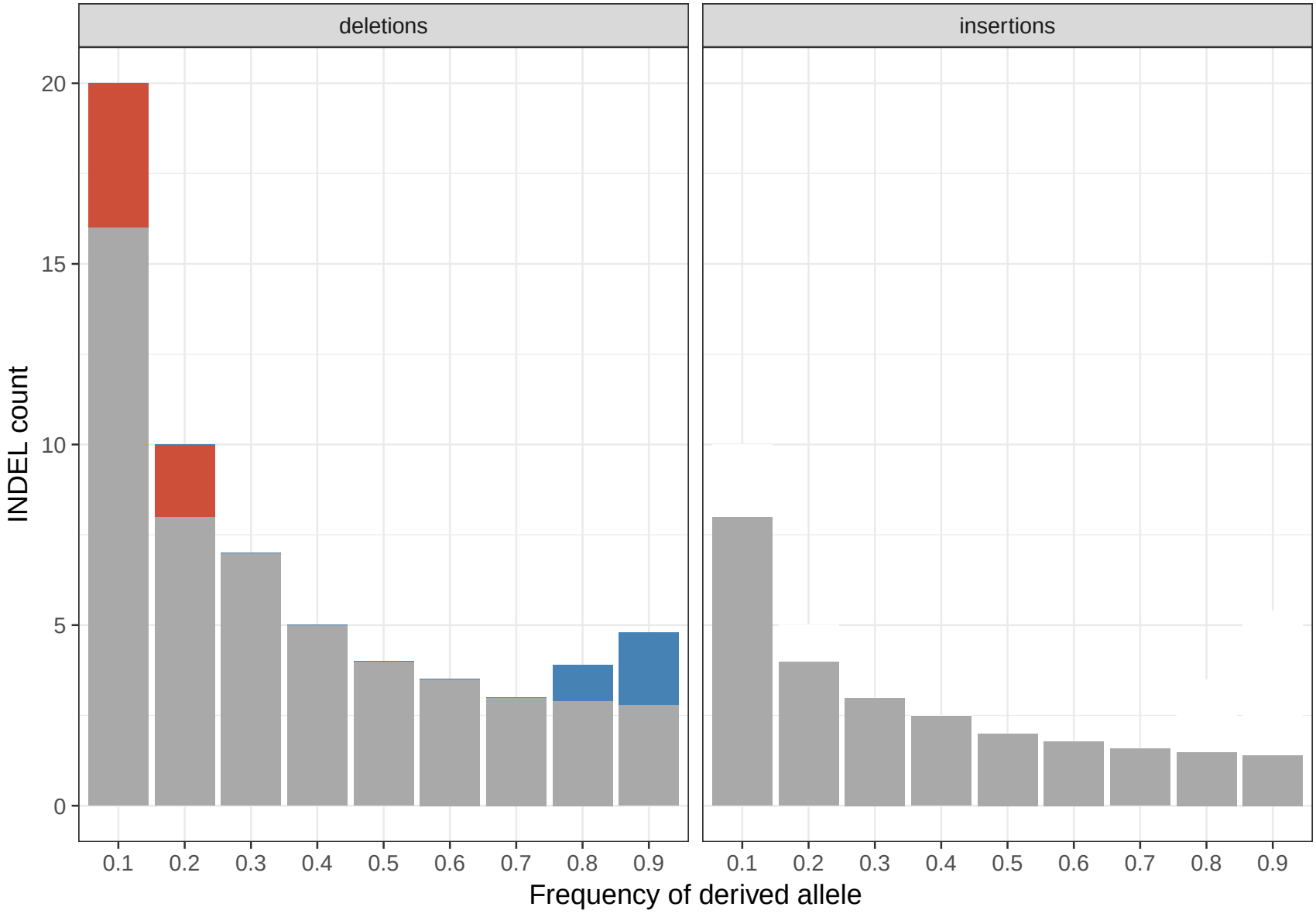
(see Hernandez et al., 2007)

Importance of polarisation error



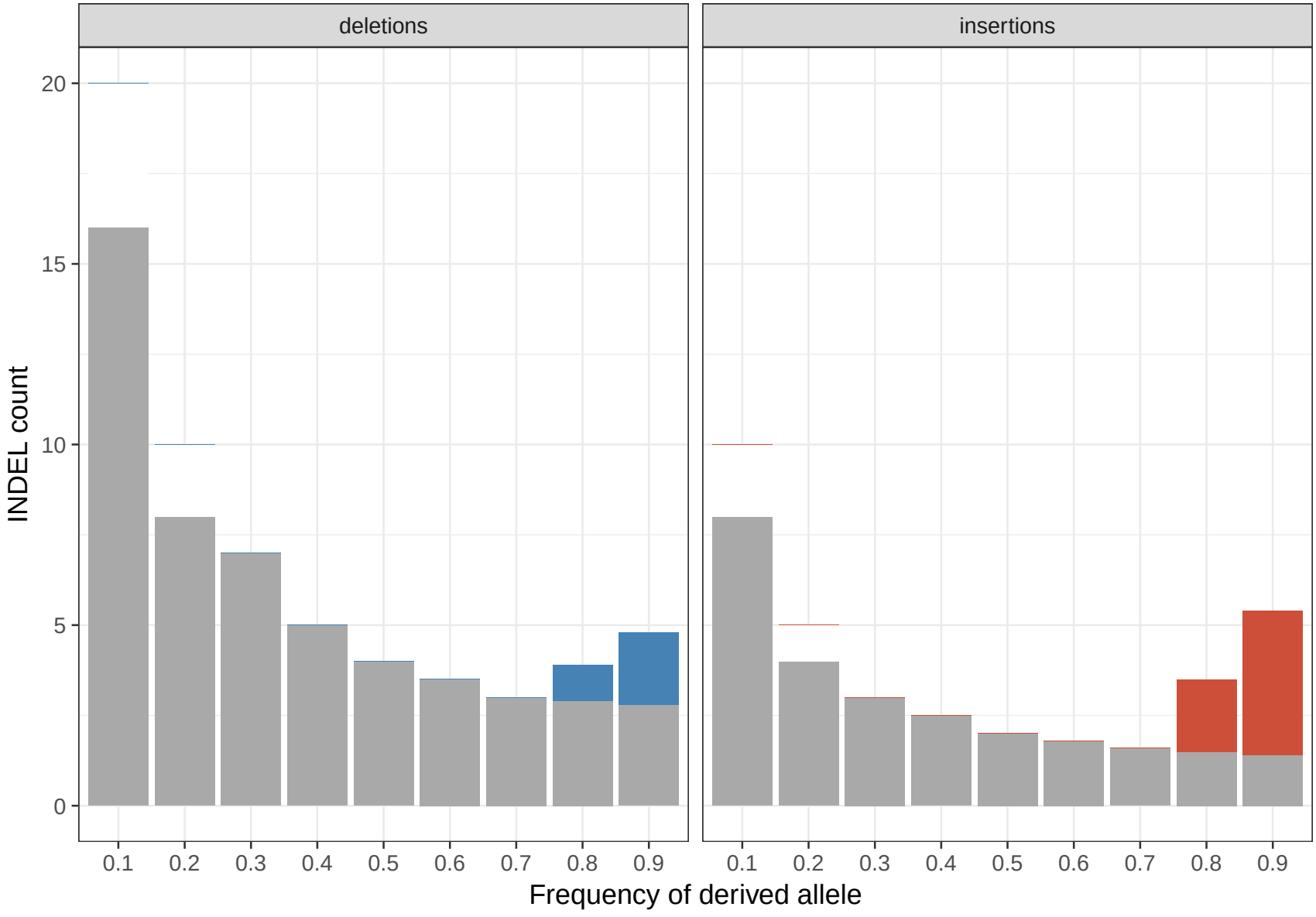
(see Hernandez et al., 2007)

Importance of polarisation error



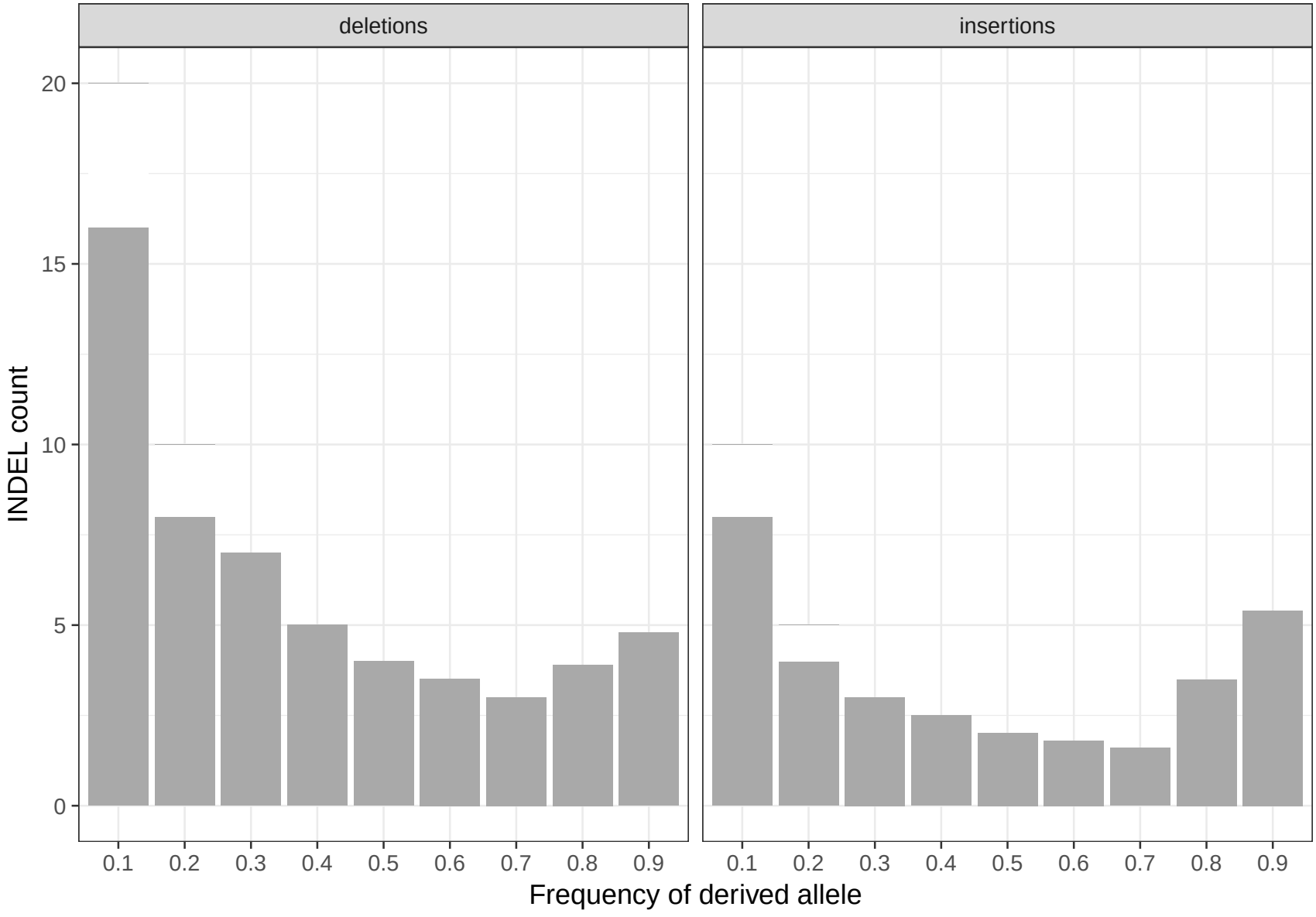
(see Hernandez et al., 2007)

Importance of polarisation error



(see Hernandez et al., 2007)

Importance of polarisation error



(see Hernandez et al., 2007)

Aims - make more specific

1. Quanitfy the selective and mutational pressures acting on INDELs in great tits (*Parus major*)
2. Investigate how these pressures vary in different genomic contexts, ie coding, non-coding, recombination rate.

Figure 2: tit

Advantages of an avian system

- Conserved karyotype and synteny - good for alignments
- Genomes consist of few large macrochromosomes and many small microchromosomes
- Results in a highly dynamic recombination landscape - power to associations with recombination

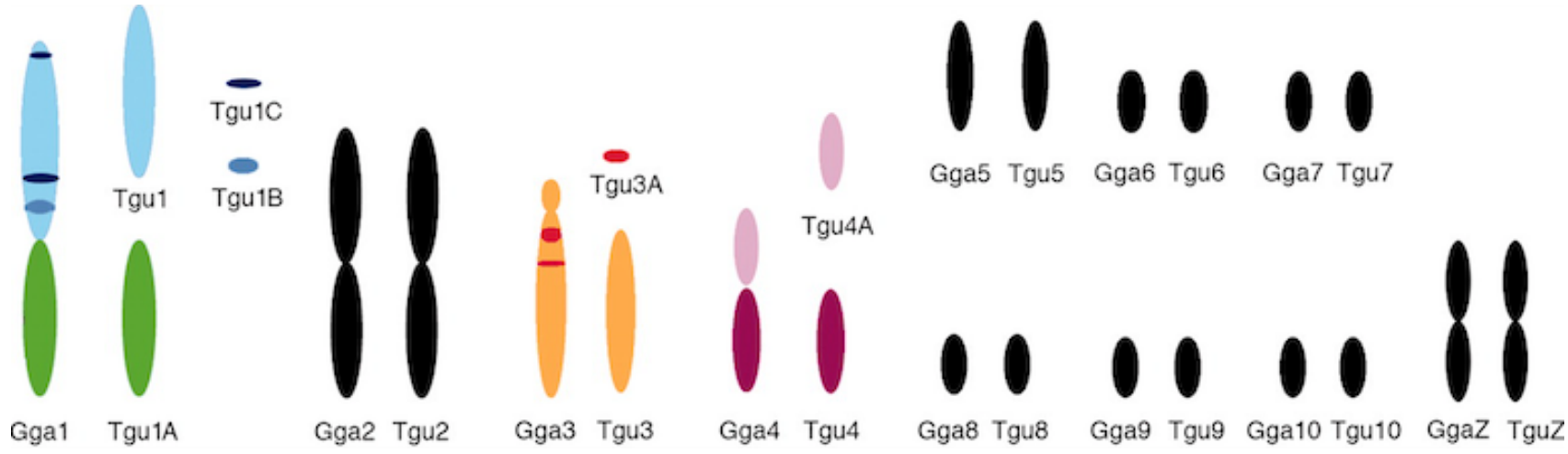


Figure 3: chroms

Data

Sample and pipeline

- 10 european great tit males (Corcoran et al., 2017)
- high coverage (44x)
- variant calling with GATK
- multispecies alignment between zebra finch, flycatcher and great tit
- parsimony based polarisation

The model - ‘anavar’

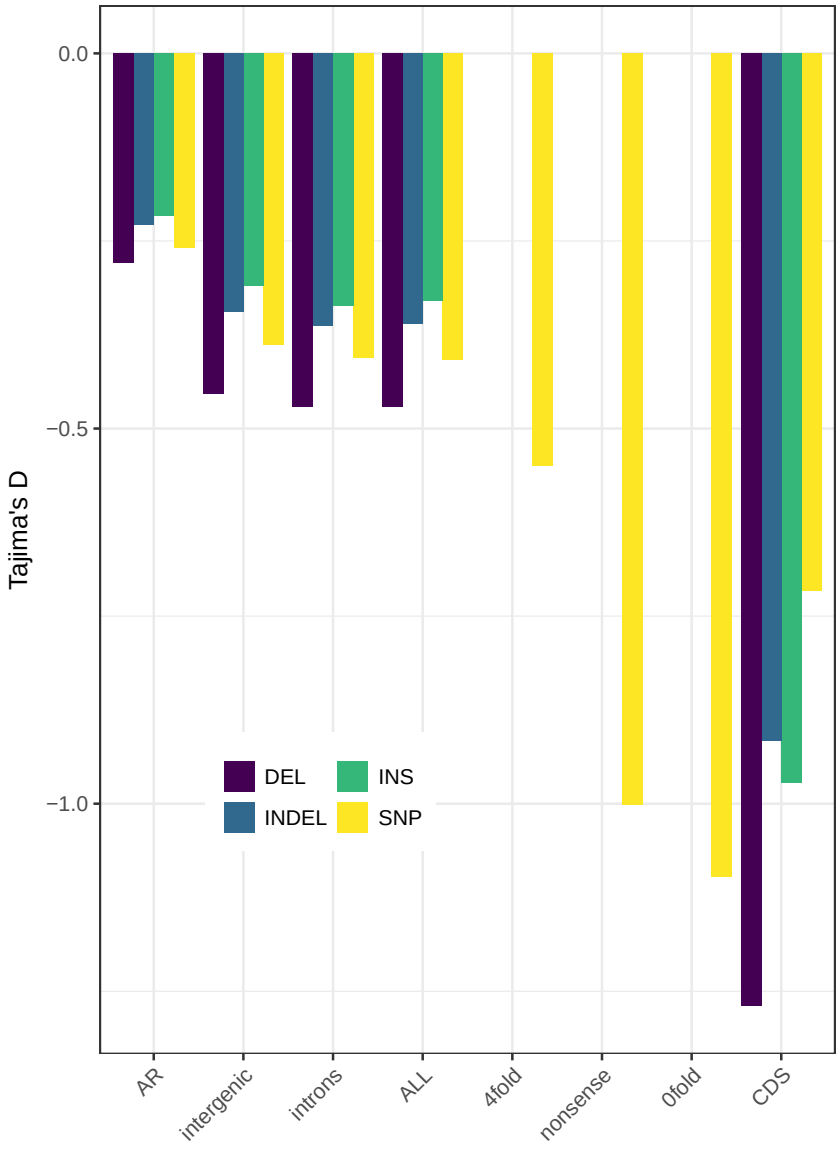
A novel maximum likelihood approach

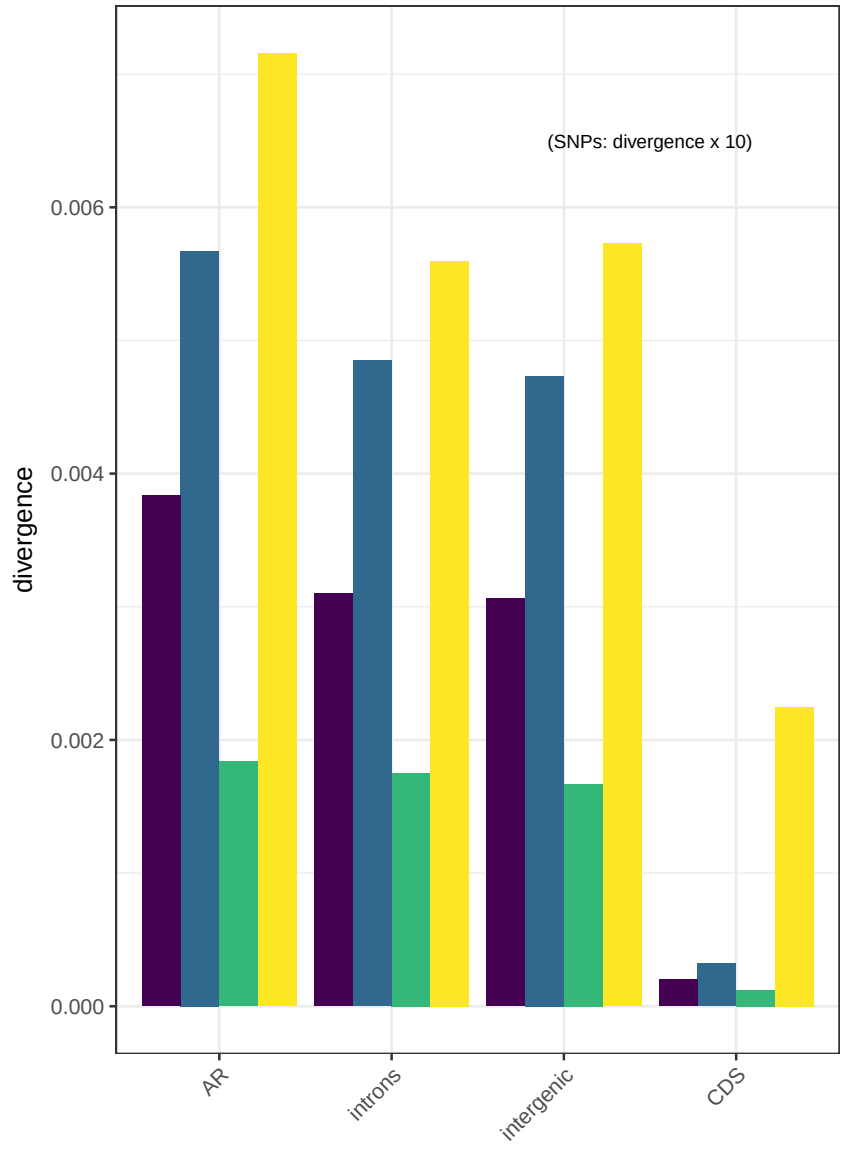
- takes the unfolded site frequency spectrum
- estimates for both insertions and deletions:
  - mutation rate ($\theta = 4N_e\mu$)
  - selection:
    - either selection coefficient ($\gamma = 4N_e s$)
    - or scale and shape paramater for distribution of fitness effects
  - polarisation error
- Controls for demography using neutral sites (Eyre-Walker et al., 2006)

(Barton and Zeng, 2018)

Dataset summary

Regional variation in purifying selection





Coding INDELs

Most INDELs extremely deleterious in coding regions

Moving away from genes

Approach

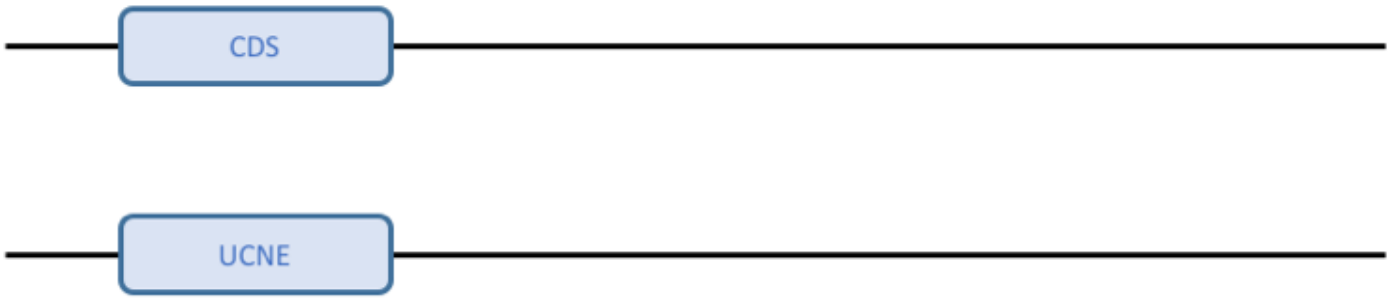


Figure 4: ls1

Approach

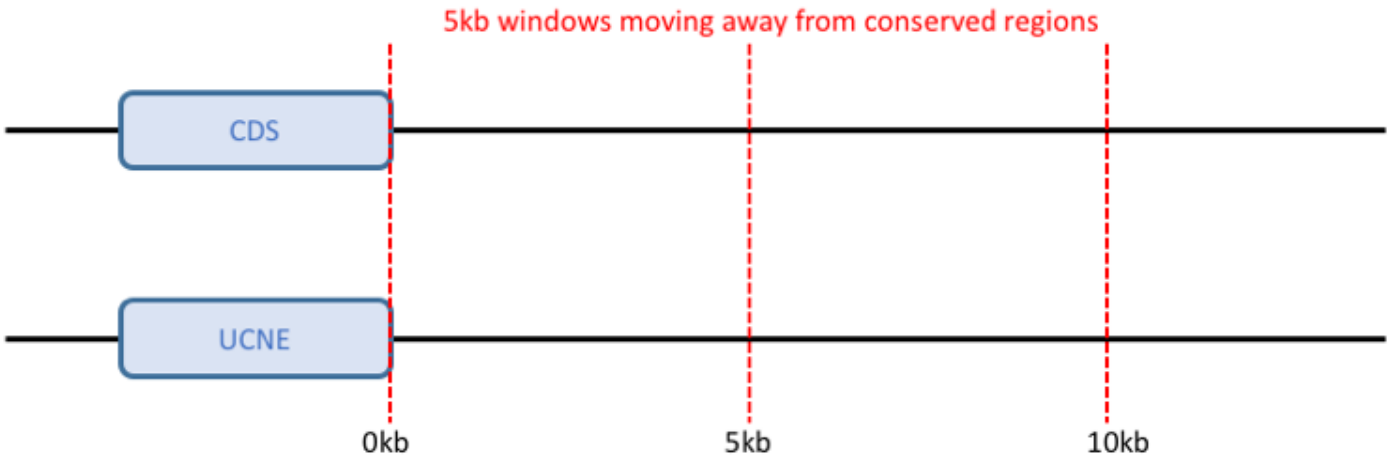


Figure 5: Is2

Approach

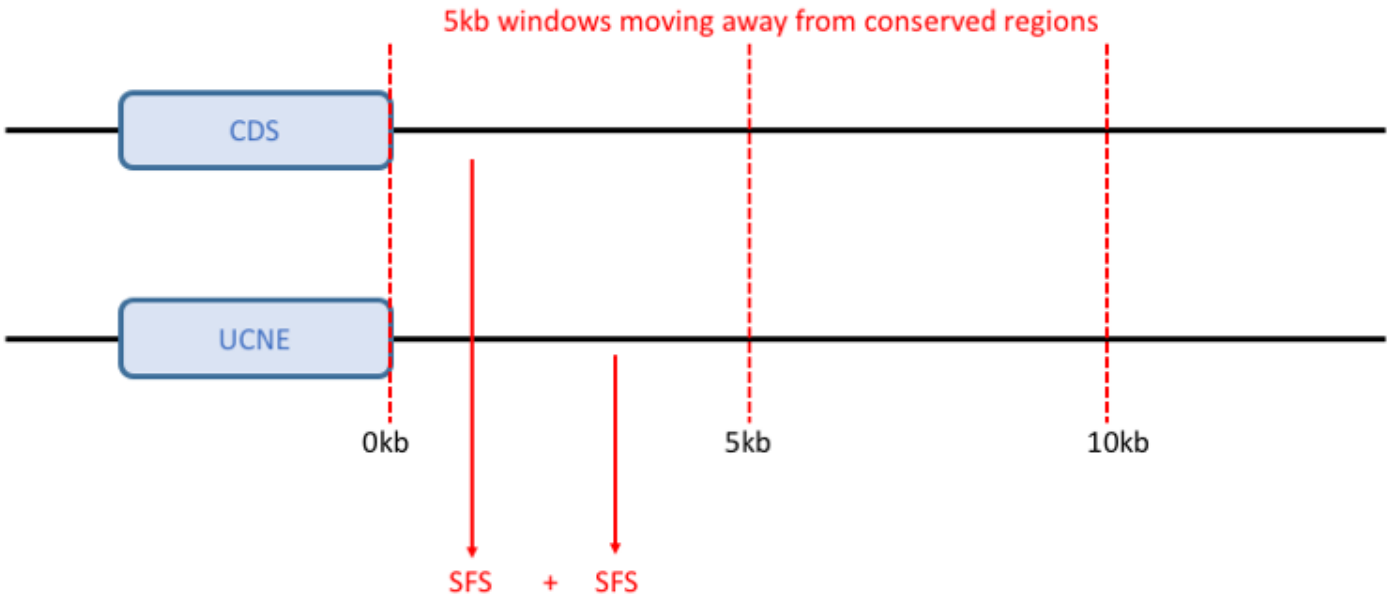


Figure 6: Is3

Approach

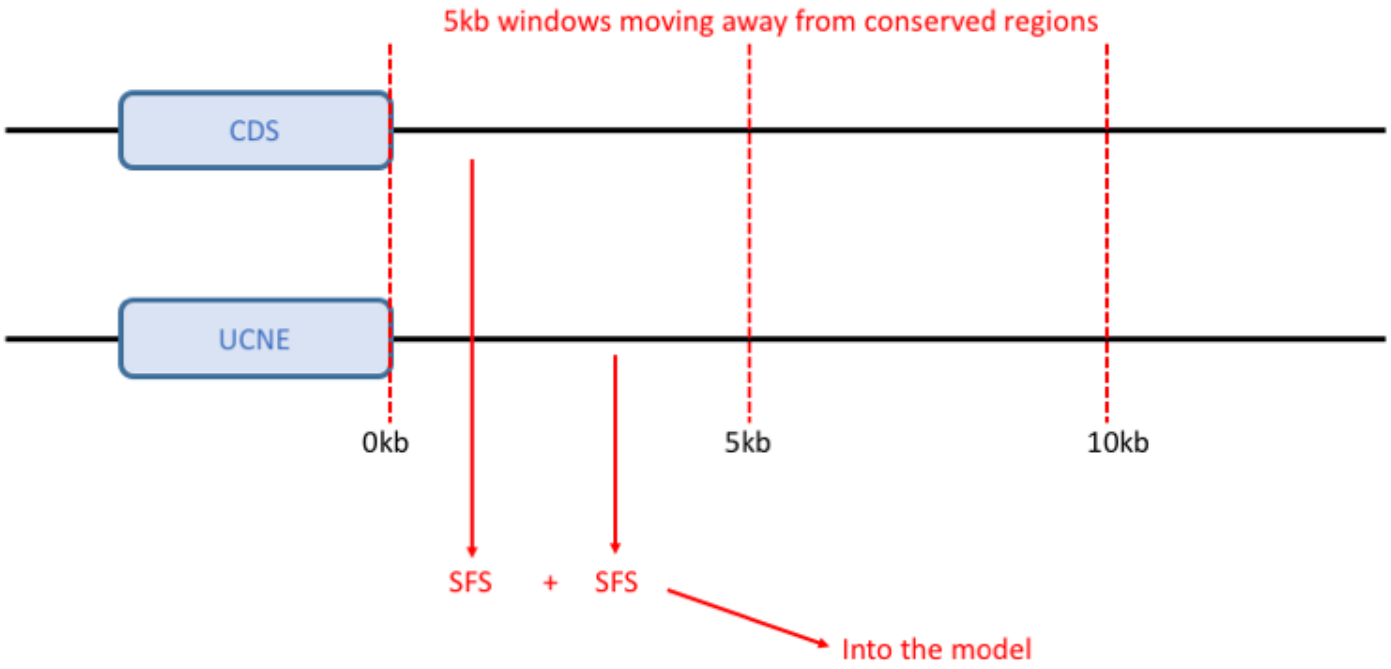


Figure 7: Is4

Evidence for linked selection

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(ins_theta$ins_theta) and ins_theta$dis
## S = 14906, p-value = 0.04581
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2842257

## [1] 50

##
##  Spearman's rank correlation rho
##
## data:  as.numeric(del_theta$del_theta) and del_theta$dis
## S = 10950, p-value = 0.0005822
## alternative hypothesis: true rho is not equal to 0
```

Recombination Analyses

Getting the data

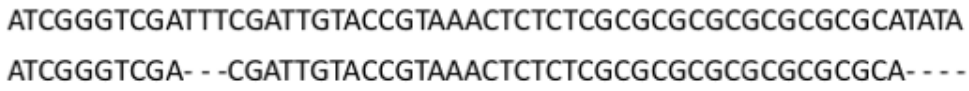


Figure 8: r1

Getting the data

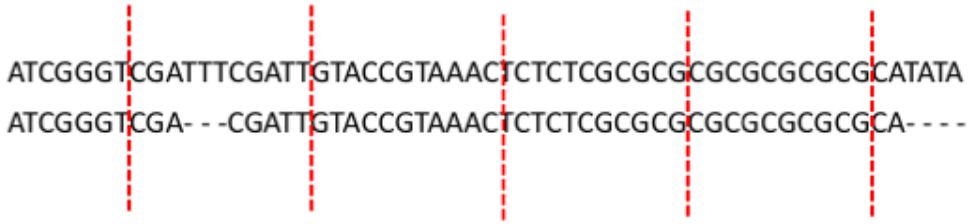


Figure 9: r2

Getting the data

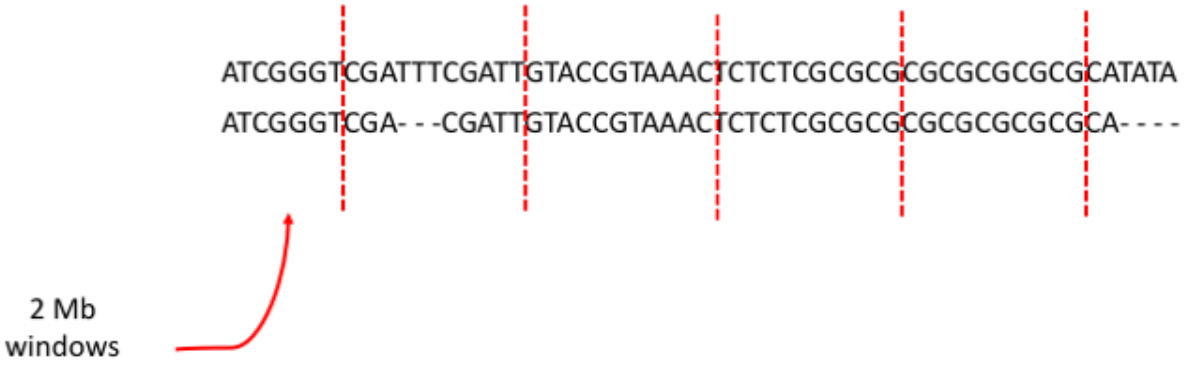


Figure 10: r3

Getting the data

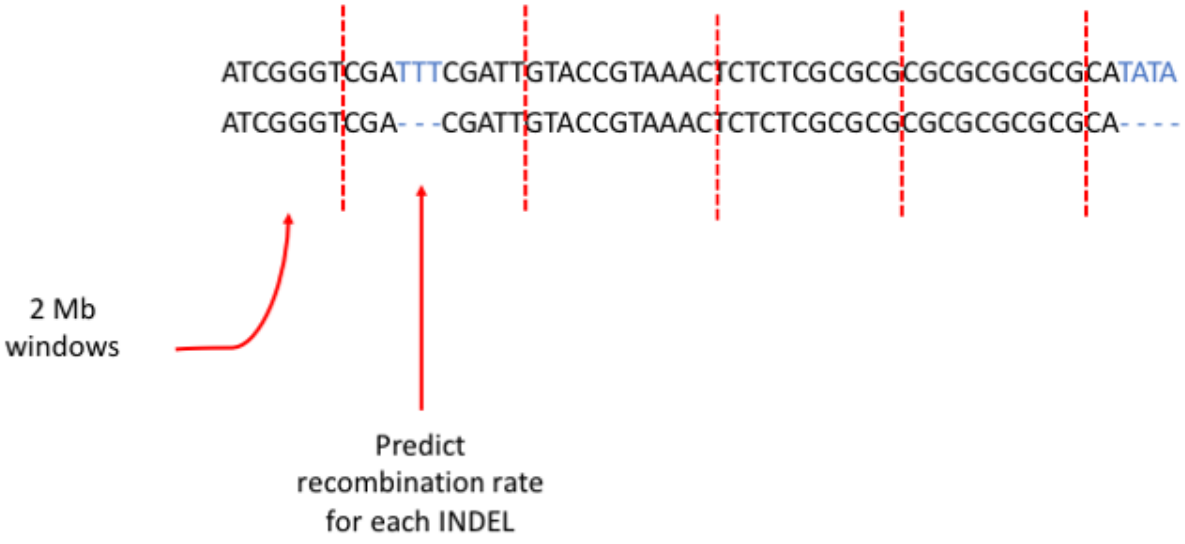
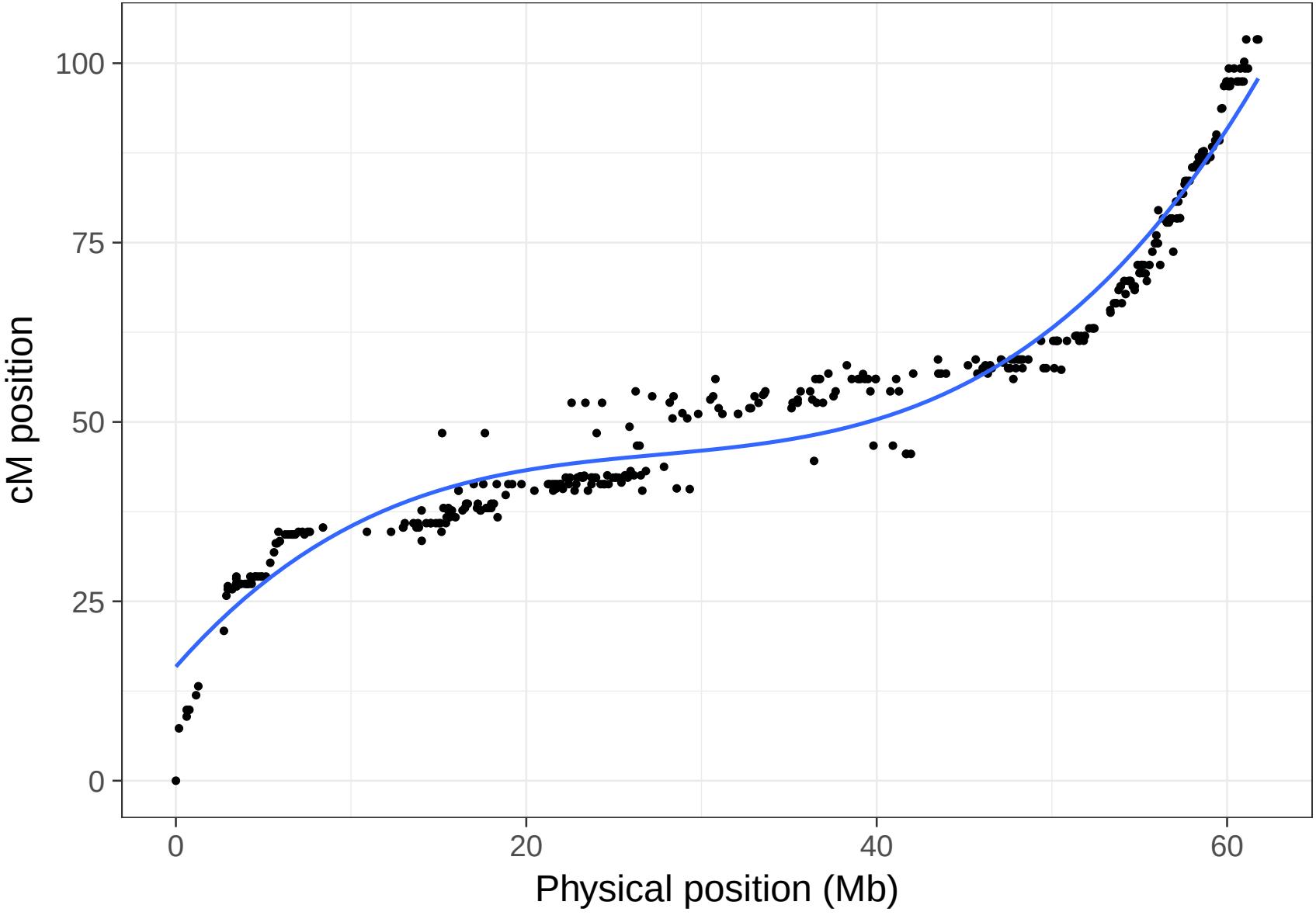


Figure 11: r4

Getting the data - recombination rate



Getting the data

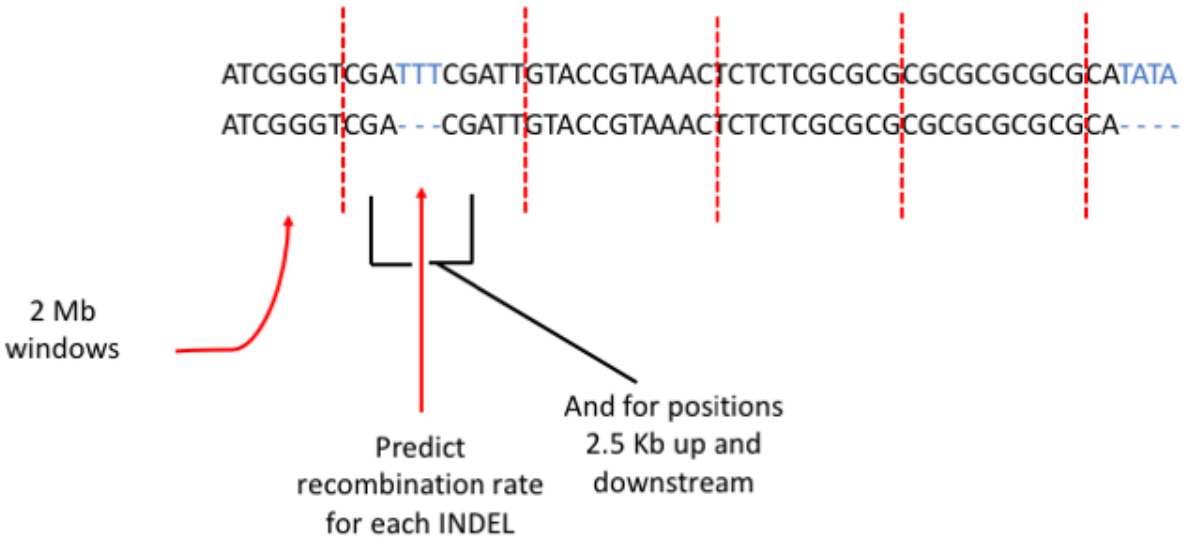


Figure 12: r5

Getting the data

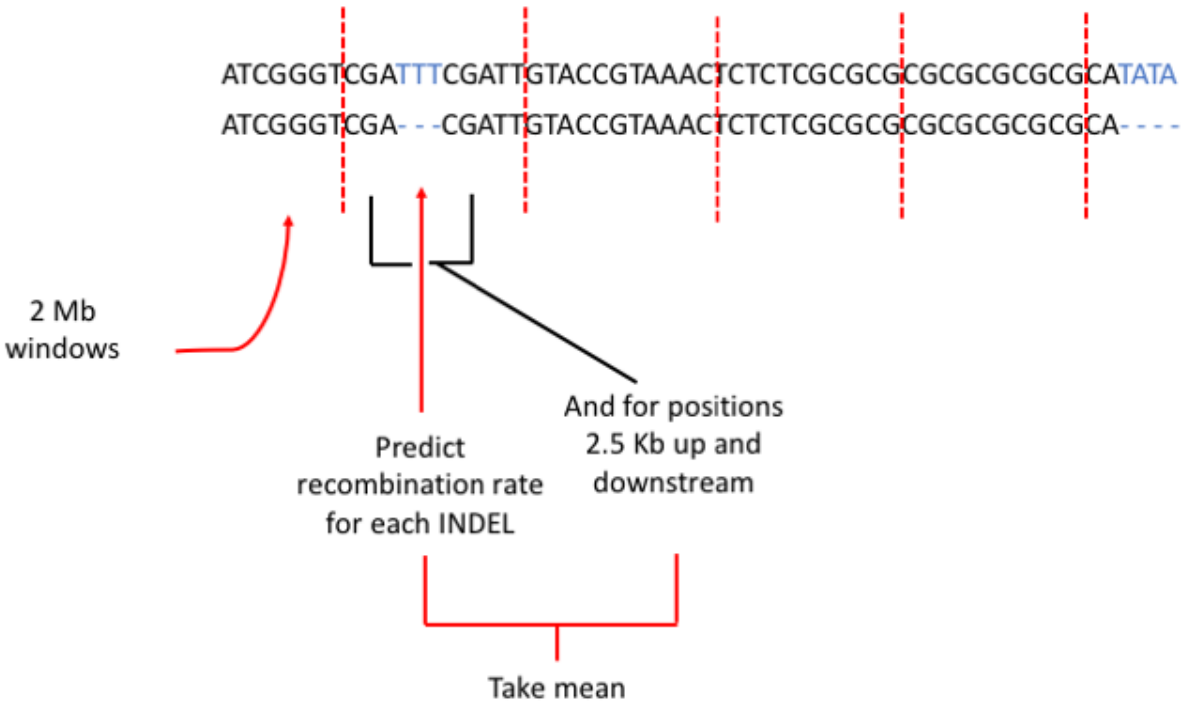


Figure 13: r6

Getting the data

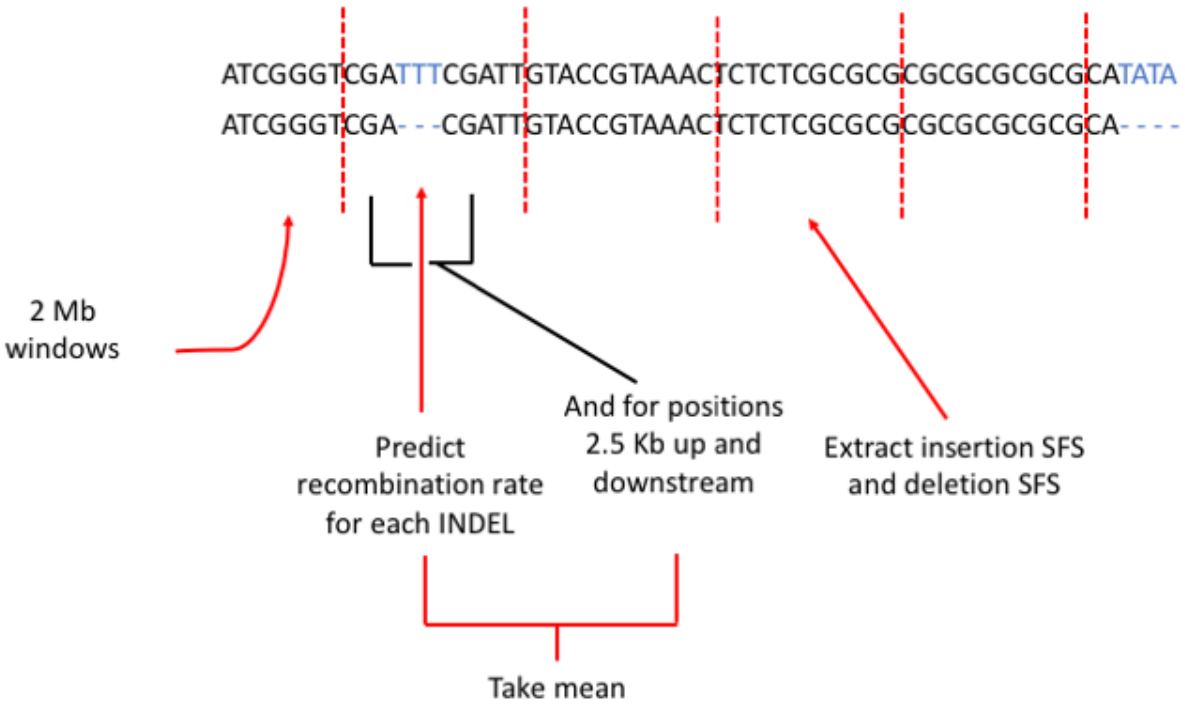


Figure 14: r7

Getting the data

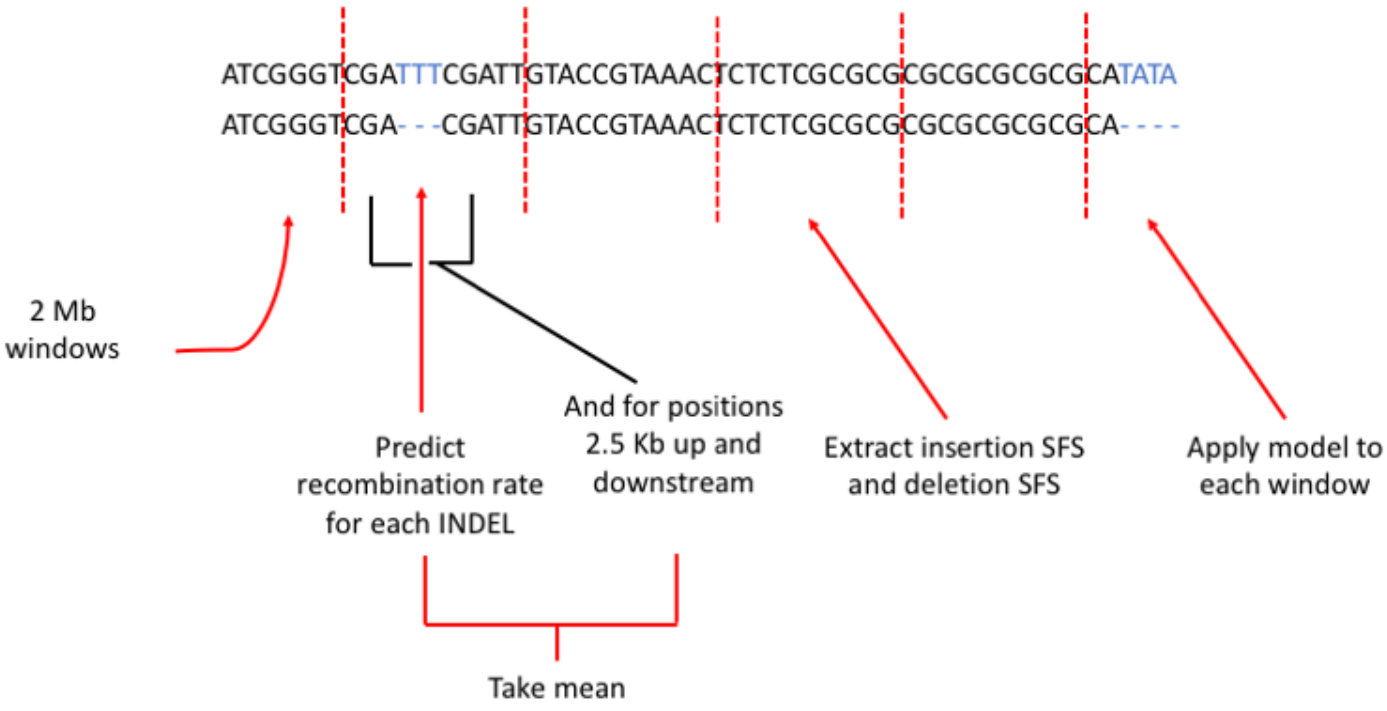


Figure 15: r8

Association between diversity and recombination

```
##
##  Spearman's rank correlation rho
##
## data:  window_data$tajd_ins and window_data$rec_rate
## S = 3969000, p-value = 3.725e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2998399

##
##  Spearman's rank correlation rho
##
## data:  window_data$tajd_del and window_data$rec_rate
## S = 3810100, p-value = 1.478e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
```

Round up

Conclusion

- INDELs in genes are bad
- Don't find a link between recombination rate and deletion bias
- Linked selection appears to be constraining INDEL rates

Next steps

- Calculate alpha - proportion of substitutions fixed by positive selection
- Separate UCNE and CDS in linked selection analysis
- Any suggestions?

Questions?